

Вероятностное тематическое моделирование

Константин Воронцов

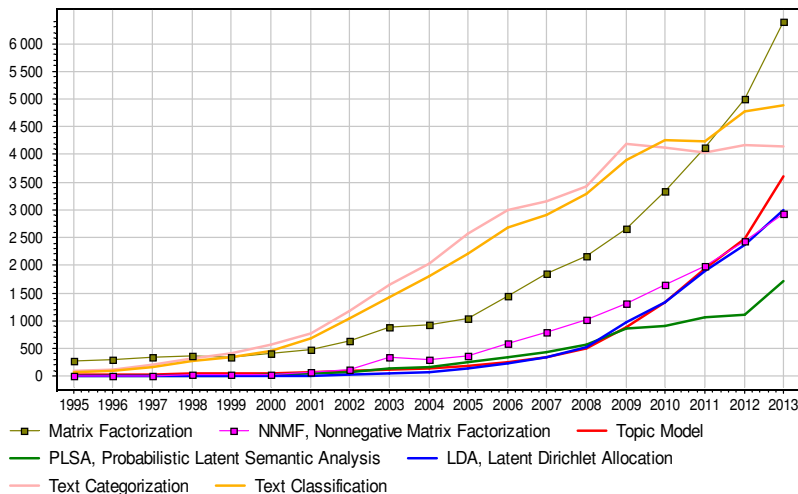
Яндекс • МФТИ • ВШЭ • МГУ • ВЦ РАН • FORECSYS

научный семинар • ШАД Яндекс • 30 сентября 2014

- 1 **Вероятностное тематическое моделирование**
 - Задача тематического моделирования
 - Модель PLSA и EM-алгоритм
 - Модель LDA
- 2 **Аддитивная регуляризация тематических моделей**
 - Проблема неединственности и неустойчивости решения
 - ARTM — задача многокритериальной оптимизации
 - Примеры регуляризаторов
- 3 **Приложения тематического моделирования**
 - Мультимодальные тематические модели
 - Прикладные задачи
 - Литература

Тематическое моделирование и близкие области исследований

Динамика цитирования, по данным Google Scholar:



Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах:
 $p(w|t)$ — вероятность встретить термин w в теме t .

Документ имеет ненаблюдаемый *тематический профиль*:

$p(t|d)$ — неизвестная частота темы t в документе d .

Когда автор писал термин w в документ d , он думал о теме t .

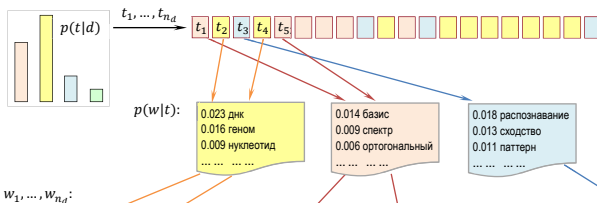
Документ d состоит из наблюдаемых терминов w_1, \dots, w_{n_d} ,

$p(w|d)$ — известная частота термина w в документе d .

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D :

$$p(w|d) = \sum_t p(w|t)p(t|d), \quad d \in D$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в **геномных** последовательностях. Метод основан на разномасштабном оценивании **сходства нуклеотидных** последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим **ортогональным базисам**. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое **распознавание повторов** различных видов (прямых и инвертированных, а также **тандемных**) на **спектральной матрице сходства**. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы **сегментных дупликаций** и **мегасателлитные участки** в **геноме**, районы **синтезии** при сравнении пары **геномов**. Его можно использовать для детального изучения фрагментов **хромосом** (поиска размытых участков с умеренной длиной повторяющегося **паттерна**).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дана коллекция текстовых документов (мешков слов):

n_{dw} — сколько раз термин w встречается в документе d

Найти модель $p(w|d) = \sum_t \phi_{wt} \theta_{td}$ с параметрами ϕ , θ :

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации логарифма правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$$

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Точка максимума $\mathcal{L}(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t|d, w)$, n_{wt} , n_{td} :

$$\begin{aligned} \text{Е-шаг:} & \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \text{М-шаг:} & \quad \begin{cases} \phi_{wt} = \frac{n_{wt}}{n_t}; & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; & n_t = \sum_{w \in W} n_{wt} \\ \theta_{td} = \frac{n_{td}}{n_d}; & n_{td} = \sum_{w \in W} n_{dw} p_{tdw}; & n_d = \sum_{t \in T} n_{td} \end{cases} \end{aligned}$$

ЕМ-алгоритм — чередование Е- и М-шага до сходимости, т. е. решение системы уравнений методом простых итераций.

✓ *Идея на будущее: можно использовать и другие методы!*

LDA — Latent Dirichlet Allocation [Blei 2003]

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \quad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Различие проявляется только при малых n_{wt} , n_{td} .

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

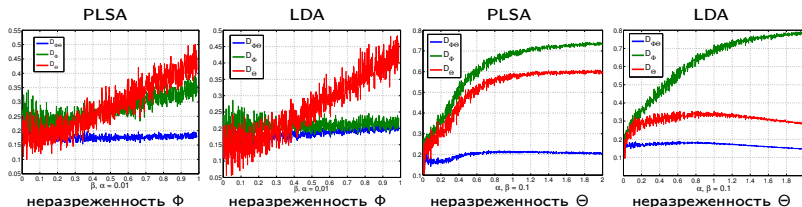
Задача построения BTM — некорректно поставленная

Неединственность стохастического матричного разложения:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для любых $S_{T \times T}$ таких, что Φ', Θ' — стохастические.

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво,
матрица Φ и матрица Θ — только когда сильно разрежены:



Вывод 1: нужны дополнительные требования к модели.

Вывод 2: требований сглаживания в LDA не достаточно.

ARTM — Аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев — регуляризаторов $R_i(\Phi, \Theta)$, $i = 1, \dots, n$.

Метод многокритериальной оптимизации — скаляризация.

Задача максимизации регуляризованного правдоподобия:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{i=1}^n \tau_i R_i(\Phi, \Theta)}_{R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geq 0; \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1$$

где $\tau_i > 0$ — коэффициенты регуляризации.

ЕМ-алгоритм с регуляризацией М-шага

Теорема

Точка максимума $\mathcal{L}(\Phi, \Theta) + R(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdw} , n_{wt} , n_{td} :

$$\begin{aligned} \text{Е-шаг:} & \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \text{М-шаг:} & \quad \begin{cases} \phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; & n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \end{cases} \end{aligned}$$

где $(x)_+ = \max(x, 0)$ — операция положительной срезки.

$$\text{PLSA:} \quad R(\Phi, \Theta) = 0$$

$$\text{LDA:} \quad R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$$

Справочные сведения. Дивергенция Кульбака–Лейблера

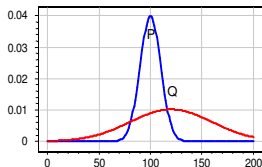
Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

$$KL(P\|Q) \equiv KL_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $KL(P\|Q) \geq 0$; $KL(P\|Q) = 0 \Leftrightarrow P = Q$;
2. Минимизация KL эквивалентна максимизации правдоподобия:

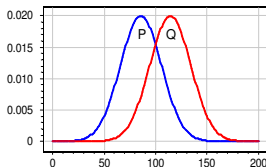
$$KL(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \rightarrow \min_{\alpha} \iff \sum_{i=1}^n p_i \ln q_i(\alpha) \rightarrow \max_{\alpha}.$$

3. Если $KL(P\|Q) < KL(Q\|P)$, то P сильнее вложено в Q , чем Q в P :



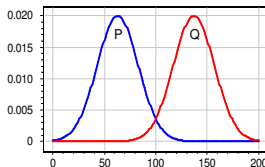
$$KL(P\|Q) = 0.442$$

$$KL(Q\|P) = 2.966$$



$$KL(P\|Q) = 0.444$$

$$KL(Q\|P) = 0.444$$



$$KL(P\|Q) = 2.969$$

$$KL(Q\|P) = 2.969$$

Примеры регуляризаторов (сглаживание и разреживание)

- ❶ разреживание предметных тем $S \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in S} \text{KL}_w(\beta_w \| \phi_{wt}) + \alpha_0 \sum_{d \in D} \text{KL}_t(\alpha_t \| \theta_{td}) \rightarrow \max$$

- ❷ сглаживание фоновых тем $B \subset T$, аналог LDA:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in B} \text{KL}_w(\beta_w \| \phi_{wt}) - \alpha_0 \sum_{d \in D} \text{KL}_t(\alpha_t \| \theta_{td}) \rightarrow \max$$

- ❸ частичное обучение по подмножествам $W_t \subset W$, $T_d \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta_{td} \rightarrow \max$$

Примеры регуляризаторов (корреляции и декорреляции)

- 4 декоррелирование тем как столбцов Φ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max$$

- 5 максимизация когерентности тем:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{w \in W} \left(\sum_{u \in W} C_{uw} n_{ut} \right) \ln \phi_{wt} \rightarrow \max$$

- 6 учёт связей между документами $n_{dd'}$:

$$R(\Theta) = \tau \sum_{d, d'} n_{dd'} \sum_{t \in T} \theta_{td} \theta_{td'} \rightarrow \max$$

- 7 учёт корреляций между темами как строками Θ :

$$R(\Theta) = -\frac{\tau}{2} \sum_{d \in D} (\ln \theta_{td} - \mu)^T \Sigma^{-1} (\ln \theta_{td} - \mu) \rightarrow \max$$

Примеры регуляризаторов (определение числа тем)

- 8 удаление неинформативных тем:

$$R(\Theta) = \tau \text{KL}_t\left(\frac{1}{|T|} \parallel p(t)\right) \rightarrow \max, \quad p(t) = \sum_{d \in D} \theta_{td} p(d)$$

- 9 разреживание тем во времени:

$$R(\Theta) = \tau \sum_{y \in Y} \text{KL}_t\left(\frac{1}{|T|} \parallel p(t|y)\right) \rightarrow \max, \quad p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$$

- 10 сглаживание тем во времени:

$$R(\Theta) = -\tau \sum_{y \in Y} \sum_{t \in T} |p(t|y) - p(t|y-1)| \rightarrow \max$$

Примеры регуляризаторов (классификация)

- 11 классификация документов по классам $c \in C$, $\psi_{ct} = p(c|t)$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \rightarrow \max$$

- 12 категоризация документов по классам $c \in C$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \sum_{t \in T} \psi_{ct} \theta_{td} \rightarrow \max$$

- 13 оптимизация AUC (D_c — множество документов класса c):

$$R(\Psi, \Theta) = -\tau \sum_{c \in C} \sum_{d \in D_c} \sum_{d' \notin D_c} \mathcal{L} \left(\sum_{t \in T} \psi_{ct} (\theta_{td} - \theta_{td'}) \right) \rightarrow \max$$

- 14 мультимодальная классификация документов:

$$R(\Phi, \Theta) = \sum_{j=1}^m \tau_j \sum_{d \in D} \sum_{x \in X_j} n_{dx} \ln \sum_{t \in T} \phi_{xt} \theta_{td} \rightarrow \max$$

Примеры модальностей в текстах

- слова — основная модальность
- слова каждого языка — отдельная модальность
- пользователи, смотревшие документ
- рекламные баннеры, просмотренные вместе с документом
- категории рубрикатора
- авторы документов
- метки времени
- документы, ссылающиеся на данный
- документы, на которые ссылается данный
- сущности (entity), упоминаемые в текстах
- признаки на изображениях, связанных с текстом

Мультимодальные тематические модели

Произвольное число модальностей X_j , $j = 1, \dots, m$.

Вероятностное пространство $D \times T \times X$, $X = X_1 \sqcup \dots \sqcup X_m$.

Каждый документ d состоит из токенов $x_1, \dots, x_{n_d} \in X$.

Тематическая модель j -й модальности:

$$p(x|d) = \sum_{t \in T} p(x|t) p(t|d) = \sum_{t \in T} \phi_{xt} \theta_{td}, \quad x \in X_j, \quad d \in D$$

Задача максимизации взвешенного правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{j=1}^m \tau_j \sum_{d \in D} \sum_{x \in X_j} n_{dx} \ln \sum_{t \in T} \phi_{xt} \theta_{td} \rightarrow \max,$$

при ограничениях нормировки и неотрицательности

$$\phi_{xt} \geq 0; \quad \sum_{x \in X_j} \phi_{xt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1$$

Модифицированный EM-алгоритм

Теорема

Точка максимума $\mathcal{L}(\Phi, \Theta) + R(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdx} , n_{xt} , n_{tdj} :

$$\text{Е-шаг: } p_{tdx} = \frac{\phi_{xt}\theta_{td}}{\sum_{s \in T} \phi_{xs}\theta_{sd}};$$

$$\text{М-шаг: } \phi_{xt} \propto \left(n_{xt} + \phi_{xt} \frac{\partial R}{\partial \phi_{xt}} \right)_+; \quad n_{xt} = \sum_{d \in D} n_{dx} p_{tdx};$$

$$\theta_{td} \propto \left(\sum_{j=1}^m \tau_j n_{tdj} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_{tdj} = \sum_{x \in X_j} n_{dx} p_{tdx}.$$

Следующие доклады

- *Анна Потапенко.* Сглаживание, разреживание и декоррелирование тематических моделей
- *Мурат Апишев.* Открытая библиотека тематического моделирования BigARTM
- *Марина Дударенко.* Мультязычные тематические модели
- *Никита Дойков.* Динамические тематические модели
- *Надежда Чиркова.* Иерархические тематические модели научных конференций ММРО и ИОИ
- *Андрей Шапулин.* Тематические модели классификации для диагностики заболеваний по электрокардиограмме

Литература

- *Hofmann T.* Probabilistic Latent Semantic Indexing // SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.
- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. — Т. 455., № 3. С. 268–271.
- *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks and Texts. Ekaterinburg, 10–12 April 2014. Springer.
- *Vorontsov K. V., Potapenko A. A.* Additive Regularization of Topic Models // Machine Learning Journal. Springer (to appear).

Воронцов Константин Вячеславович
voron@yandex-team.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели
(курс лекций, К. В. Воронцов)
- Тематическое моделирование